Article

Patterns of Ancestral Animal Codon Usage Bias Revealed Through Holozoan Protists

Jade Southworth, Paul Armitage, Brandon Fallon, Holly Dawson, Jarosław Bryk & Martin Carr*

Department of Biological & Geographical Sciences, University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom

*Author for correspondence: Martin Carr, School of Applied Sciences, University of Huddersfield, Huddersfield, West Yorkshire, United Kingdom, Telephone: +44 484-471608, email: M.Carr@hud.ac.uk

Abstract

Choanoflagellates and filastereans are the closest known single celled relatives of Metazoa within Holozoa and provide insight into how animals evolved from their unicellular ancestors. Codon usage bias has been extensively studied in metazoans, with both natural selection and mutation pressure playing important roles in different species. The disparate nature of metazoan codon usage patterns prevents the reconstruction of ancestral traits. However, traits conserved across holozoan protists highlight characteristics in the unicellular ancestors of Metazoa. Presented here are the patterns of codon usage in the choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta*, as well as the filasterean *Capsaspora owczarzaki*. Codon usage is shown to be remarkably conserved. Highly biased genes preferentially use GC-ending codons, however there is limited evidence this is driven by local mutation pressure. The analyses presented provide strong evidence that natural selection, for both translational accuracy and efficiency, dominates codon usage bias in holozoan protists. In particular, the signature of selection for translational accuracy can be detected even in the most weakly biased genes. Biased codon usage is shown to have co-evolved with the tRNA species, with optimal codons showing complementary binding to the highest copy number tRNA genes. Furthermore, tRNA modification is shown to be a common feature for amino acids with higher levels of degeneracy and highly biased genes show a strong preference for using modified tRNAs in translation. The translationally optimal codons defined here will be of benefit to future transgenics work in holozoan protists, as their use should maximise protein yields from edited transgenes.

Keywords: Choanoflagellates, *Capsaspora*, optimal codons, premetazoan, translational accuracy, tRNA modification

Background

The closest known relatives of metazoans consist of multiple lineages of unicellular eukaryotes known collectively as the holozoan protists. Within Holozoa the choanoflagellates are the sister-group to Metazoa (Carr et al. 2008), with the Filasterea being recovered as a more distantly related lineage (Shalchian-Tabrizi et al. 2008). The choanoflagellates are a large, diverse group of aquatic filter feeders, found in both freshwater and marine environments (see Leadbeater 2015 for a thorough review on the group). In contrast, only five species of filasterean have been described. Two of these are marine bacteriovores in the genus *Ministeria* (Patterson et al. 1993; Tong 1997) and two are predatory flagellates in the genus *Pigoraptor* (Hehenberger et al. 2017). The fifth species, *Capsaspora owczarzaki*, is a symbiont of the freshwater snail *Biomphalaria glabrata* (Stibbs et al. 1979; Hertel et al. 2002).

Studying their closest relatives has shone new light on the evolution of metazoans (Carr et al. 2010, 2017; Suga et al. 2013; Tucker 2013; Najile et al. 2016) and it is now clear that the last common ancestor of the holozoans had a complex genome, containing genes previously thought to be unique to Metazoa (Fairclough et al. 2013; Sebé-Pedrós et al. 2016; Hehenberger et al. 2017). Traits which are shared by both choanoflagellates and filastereans are candidates for being ancestral to the two groups. By extension such traits are also ancestral to Metazoa, even if the traits are not present in extant metazoans. Such traits must have been lost, either in a unicellular premetazoan or stem-group metazoan (Figure 1).

Whole genome sequences from the choanoflagellates *Monosiga brevicollis* (King et al. 2008) and *Salpingoeca rosetta* (Fairclough et al. 2013), as well as *C. owczarzaki* (Suga et al. 2013), have highlighted extensive gene loss and gain within holozoans; however, such studies have not examined the genomes on a population genetics level in order to determine the role of natural selection in their evolution. One study that has considered aspects of the population biology of choanoflagellates was that of Carr et al. (2017), which analysed transcriptome sequences of 19 species of choanoflagellate to highlight variation in the strength of natural selection in the elongation factor genes EF-1A and EFL. The study used rates of nucleotide substitutions at synonymous and non-synonymous sites (Ka/Ks) and also codon usage analyses to examine varying selective constraint across species.

The genetic code shows redundancy and Clarke (1970) speculated that there may be numerous selective forces controlling the codon preference employed by genes. Grantham et al. (1980) showed that codons are indeed used in a non-random fashion within eukaryotes, prokaryotes and viruses. The level of codon usage bias in a gene is commonly determined using the 'effective number of codons' or *Nc* (Wright 1990). Values of *Nc* range from 20, in genes where each amino acid only uses a single codon, to 61, when codons are used equally for each amino acid. Numerous subsequent studies have shown that both natural selection and mutational pressures can shape codon usage (Sharp et al. 1995; Smith and Eyre-Walker 2001; Figuet et al. 2014). For those species where codon usage is driven by natural selection, it is assumed that they possess a suite of translationally 'optimal codons' that provide a selective advantage when used in preference to non-optimal codons. As originally defined (Ikemura 1981), optimal codons are those codons that complement the most abundant isoaccepting tRNA. Lloyd and Sharp (1991, 1993) proposed an alternative

93 strategy to determine optimal codons, identifying those codons whose frequency of
94 usage is significantly higher in putatively highly expressed genes. A third
95 methodology, as implemented by the program CodonW (Peden, 1999) when
96 expression data are absent, is to identify those codons are at significantly higher
97 frequencies in the most biased genes, based upon axis 1 of a correspondence analysis,
98 compared to the least biased genes in a genome. Peden (1999) noted that codons
99 identified as optimal by CodonW can be considered to be genuine if the major trend
100 in codon usage is for optimal translation and that highly biased genes are highly
101 expressed. The usage of optimal codons in a gene can be described using F_{op}, the
102 Frequency of Optimal Codons, (Ikemura 1981), calculated by dividing the number of
103 optimal codons present in a gene by the total number of codons.

104 Selection on the use of optimal codons may operate through two, not mutually
105 exclusive, mechanisms. Codon usage bias is often more extreme in highly expressed
106 genes (Gouy and Gautier 1982) and optimal codons may facilitate more rapid
107 translation than non-optimal codons (Pedersen 1984; Sørensen and Pedersen 1991).
108 This led to the view that selection for optimal codons may be a result of selection for
109 translational efficiency (Ehrenberg and Kurland 1984). In unicellular organisms, the
110 rate of protein synthesis can have a profound effect on growth rate (Bulmer 1991) and
111 therefore reproductive success, highlighting the importance of selection upon
112 translational efficiency in protists. Proteins that contain misincorporated amino acids
113 will constitute a metabolic burden and therefore selection upon codon usage may also
114 reflect the requirement for accurately translated polypeptides. Consistent with this
115 view, Precup and Parker (1987) showed that favoured codons in *Escherichia coli*
116 could lead to a 10-fold reduction in amino acid misincorporation compared to non-
117 favoured codons. Akashi (1994) provided further evidence for selection acting upon

translational accuracy by showing that preferred codons were enriched in regions encoding DNA-binding domains compared to non-domain regions in *Drosophila* transcription factors.

Selection coefficients (s) associated with the use of optimal codons over non-optimal ones are estimated to be small, with s=1x10^{-6}-1x10^{-8} in *Drosophila* (Hartl et al. 1994; Akashi 1995; Maside et al. 2004). Therefore, for selection to operate on synonymous codons, a species must possess a large effective population size (N_e). Effective population sizes are currently unknown for choanoflagellate and filasterean species. Evidence from other unicellular eukaryotes indicates a broad range of values of N_e, with effective population sizes in some taxa being similar to multicellular organisms but 2-4 orders of magnitude higher in other species (Snoke et al. 2006; Watts et al. 2013). Lynch and Conery (2003) argued that the transition from a unicellular existence to multicellularity in eukaryotes resulted in order-of-magnitude reductions in N_e, indicating that selection could play a stronger role in codon usage bias in holozoan protists than in metazoans.

In the absence of efficient natural selection, codon usage is likely to be determined by mutational forces. Mutation pressures which are strong enough to bias neutral synonymous codon positions are also likely to influence nucleotide composition at non-coding sites such as introns and intergenic regions. A strong relationship between the GC-content at silent third positions (GC3s) and local base composition may indicate that mutational pressure is a major driver of codon usage; the absence of such a relationship would suggest that mutation does not play a strong role and points to natural selection as being the major player in codon choice (Kliman and Hey 1994).

142 The Carr et al. (2017) choanoflagellate study showed that genes encoding
143 elongation factors, which are highly expressed (Liu et al. 1996), exhibited strong
144 codon usage bias. This finding is consistent with selection on codon usage within
145 choanoflagellates and promotes further investigation into the driving forces behind
146 codon usage bias in holozoan protists. Some choanoflagellate species have
147 intercontinental oceanic distributions (Carr et al. 2008, 2017; Nitsche et al. 2011),
148 raising the possibility that some choanoflagellate taxa may possess very large
149 population sizes. In contrast, *C. owczarzaki* is a symbiont, or parasite, of the snail
150 *Biomphalaria glabrata* and therefore may undergo regular bottlenecks as it passes
151 from one host to another; the result of such bottlenecks may act to suppress the
152 effective population size of *C. owczarzaki*. A prediction of this symbiotic lifestyle is
153 that *C. owczarzaki* may show less efficient selection operating upon its codon usage
154 compared to free-living choanoflagellates.

155 The study presented here aims to investigate the forces that control codon
156 usage bias within the three holozoan protists that currently have available whole
157 genome sequences. General trends of codon usage and optimal codons identified have
158 been determined for each species. Furthermore, the roles of mutational pressure and
159 selection, on both translational efficiency and accuracy, were considered.

160 **Results**

161 **Strength and Direction of Codon Usage Bias In The Holozoan Protists**

162 The degree of codon usage bias was determined for all genes in the
163 transcriptomes of *M. brevicollis*, *S. rosetta* and *C. owczarzaki* using the effective
164 number of codons (*Nc*). Across the three species, values of *Nc* ranged from 20-61 and
165 the mean transcriptome values are shown in Table 1. *S. rosetta* shows the strongest

166 codon usage bias across its transcriptome, whilst *M. brevicollis* and *C. owczarzaki*
167 have similar, and higher, mean values of *Nc*.

168 The proportion of GC at synonymous third positions (GC3s) highlights the
169 direction of bias, toward GC or AT-ending codons. Figure 2 shows genes in all three
170 species exhibit a strong GC bias, with highly biased genes showing higher GC3s.
171 Consistent with their weaker codon usage bias, mean GC3s is lower in both *M.*
172 *brevicollis* and *C. owczarzaki* than *S. rosetta* (Table 1). Genes enriched for AT-ending
173 codons are in a minority in all three species, with only 1.3%, 0.6% and 1.5% of genes
174 in *M. brevicollis*, *S. rosetta* and *C. owczarzaki* respectively showing GC3s lower than
175 0.5.

176 The *Nc* plot of *M. brevicollis* differed from those of *S. rosetta* and *C.*
177 *owczarzaki* due to the presence of approximately 200 genes which did not conform to
178 the general trend of increasing GC3s with increasing bias. These genes were
179 investigated further and most were shown to be 1) highly repetitive in their conceptual
180 amino acid sequence, 2) lacking identifiable functional domains, 3) lacking
181 identifiable orthologues through BLAST in either *S. rosetta* or *C. owczarzaki* and 4)
182 unable to be recovered when reciprocal BLASTs were performed with their top
183 BLASTn hits (Figure S1). Of the 200 genes examined, 159 could not be recovered
184 with reciprocal BLAST analyses and only 8 did not meet any of the four criteria
185 above. This evidence points to many of these genes being false positive annotations
186 rather than genuine genes. Excluding the 192 potential false positives has limited
187 impact on the overall genome statistics for *M. brevicollis* (mean *Nc* = 48.28±5.356,
188 mean GC3s = 0.641±0.057). As there is a lack of empirical evidence to confirm the

189 genes as false positives, they have been retained in subsequent analyses with the
190 proviso that they may be introducing noise into the results for *M. brevicollis*.

191 For species, the 5% highest, lowest and mid-biased genes were identified and
192 used to study patterns of codon usage and base composition across the transcriptome.
193 Within the *M. brevicollis* genome 75% of genes have been assigned KOG ontology
194 functional groups and gene categories (King et al. 2008), allowing a comparison of
195 ontology categories across codon usage bias categories (Table S1). The composition
196 of three of the four functional KOG groups differed between the high bias genes and
197 the other two bias categories. Genes involved in information storage and processing,
198 as well as metabolism were significantly enriched in the highly biased genes (Fisher's
199 exact tests, $P<0.02$), whilst there was a paucity of poorly characterized genes in the
200 high bias genes. Of the 25 KOG gene categories, four were significantly
201 underrepresented in the high codon bias genes compared to the mid and low bias
202 genes ($P<0.005$ in all four categories). Two of the underrepresented categories were
203 for either general function, or function unknown genes; however, there was also a
204 dearth of genes involved in signal transduction, as well as DNA replication and
205 recombination, in the highly biased genes ($P<0.005$). Genes involved in protein
206 translation made up the highest proportion of highly biased genes and this KOG
207 category was significantly enriched in the high bias genes ($P<0.0001$). Two further
208 categories, containing genes involved in energy production, as well as amino acid
209 metabolism, were also enriched ($P<0.02$ in both KOG categories).

210 Whilst all three holozoans exhibit a strong bias towards GC-ending codons it
211 is important to determine the evolutionary pressures that drive synonymous codons
212 towards guanine and cytosine. It is therefore also important to determine whether, if

selection is indeed operating upon synonymous codon usage, selection is operating at similar levels of efficiency across the three species.

The Role of Local Mutation Pressure in Determining Codon Usage

One possible explanation for the relationship between *Nc* and GC3s is that codon usage is driven mainly by mutation pressure. Under the mutational model, such a bias toward GC in highly biased genes is likely to affect all neutral nucleotide positions within a given gene. In order to determine if mutation pressure is driving the high GC3s observed in highly biased genes, values of mean GC3s, as well as GC content in introns and flanking DNA were determined for the genes within the three categories of bias (Table S2).

Figure 3 shows that for each species GC3s decreases from the highly biased to medium and least biased genes, with significant differences in GC3s observed between categories for all species (*t* test, $P<0.0001$ in each comparison, Table S2). A *t* test could not be performed on the high and mid categories for *M. brevicollis*, as the putative erroneously annotated genes, which have relatively lower GC3s, produced a non-normal distribution for this category (Figure S2). In contrast, GC content produces very different patterns between all bias categories for non-coding DNA (both flanking DNA and introns) in each of the three species (Table S2). Only *S. rosetta* shows decreasing GC content as bias level decreases in non-coding DNA, which would be expected if mutation pressure was driving the high GC3s observed in the high bias gene categories in all three species. Within this species however there is no significant difference in GC content between mid and low bias genes for either flanking DNA or introns, in contrast to the significant difference in GC3s. *M. brevicollis* shows the highest non-coding GC content in mid bias genes, whilst *C.*

owczarzaki shows the exact opposite pattern to GC3s, with the highest non-coding GC content observed in low bias genes and the lowest GC content recovered in the high bias genes (Figure 3, Table S2). The GC content flanking DNA and introns in each bias category show similar patterns in each species, suggesting mutation patterns are similar across individual genes. Plotting GC3s against the GC content of introns failed to show any relationship between the two statistics (graphs not shown), with R^2 values of 0.004, 0.007 and 0.018 given for *M. brevicollis*, *S. rosetta* and *C. owczarzaki* respectively.

The stop codons of all three species also fail to show evidence for a GC-bias in mutation pressure. The highly biased genes, based upon *Nc*, in each species show a preference for the GC-free UAA stop codon over both guanine containing stop codons (Table S3). Following on from these findings, it appears that a mutational pressure towards guanine and cytosine is not a major driver of the variation observed in GC3s.

Identification of Optimal Codons and Major tRNA Genes

The findings above suggest that natural selection for optimal codons is therefore an important force in determining codon choice. Optimal codons for all three species were determined using CodonW. In addition, optimal codons were also identified by comparing the 5% most highly and weakly expressed genes in *S. rosetta* and *C. owczarzaki*. The two strategies differed by two optimal codons for *S. rosetta* and one optimal codon for *C. owczarzaki*, indicating that the CodonW estimated optimal codons for *M. brevicollis* are likely to be accurate (Table 2).

Consistent with the major bias toward GC-ending codons across the genome, 63 out of the 69 CodonW-estimated optimal codons across the three species are GC-ending; despite this, each species also possesses at least one optimal codon ending in

uracil. In every case uracil ending codons are optimal for threefold, fourfold or sixfold degenerate amino acids and are also the least frequently used optimal codon for the encoded amino acid (Table S3). No optimal codons in any of the species possess adenine at synonymous sites. Eleven of the eighteen amino acids showing redundancy have identical optimal codons across all three species (Table 2). The remaining seven amino acids exhibiting redundancy all share at least one optimal codon, showing that the suite of optimal codons for all three species are remarkably similar.

tRNA genes were identified in each species, in order to compare optimal codons and major tRNA genes (either the most abundant or only tRNA gene for a given amino acid). The number and range of tRNA genes are also very similar across the three species. The number of identified tRNA genes ranges from 104-114 (Table S4), with *C. owczarzaki* harbouring a lower number than either choanoflagellate. It can also be seen that the numbers of tRNA genes for each amino acid are also very similar, with only four out of 20 amino acids varying by more than two tRNA gene copies across the three species (Table S4).

With the exception of lysine in *M. brevicollis*, which does not possess a single most abundant tRNA gene, there was a perfect match between optimal codons and major tRNA genes in all twofold-degenerate amino acids (Table 2). All three species have GGC as an optimal codon for glycine and also have the complementary GCC anticodon in their most abundant glycine tRNA genes (Table 2). In contrast to this, for the remaining eight threefold to sixfold-degenerate amino acids there were only six matches between optimal codons and major tRNA genes across the three species (Table 2). For those six matches, the optimal codons ended in uracil, however the encoded amino acids showed a greater preference for GC-ending optimal codons in

highly biased genes (Table S3). In the threefold to sixfold-degenerate amino acids, most of major tRNA genes (22 out of 24) harbour adenine at the anticodon site complementary to the synonymous codon position. This suggests that the codons may rely upon the wobble effect for binding to their corresponding tRNA molecule. However, deamination of the adenine base in the anticodon, at the wobble position, to inosine will allow complementary base pairing to the cytosine nucleotides of optimal codons.

In the higher degeneracy amino acids, there is a complete absence of tRNA genes that perfectly complement cytosine-ending optimal codons (Table S5). Direct evidence for tRNA modification and deamination of adenine at the wobble position to inosine was provided by screening publicly available SRA transcriptome datasets. During reverse transcription inosine in modified RNA is replaced with guanine in cDNA molecules (Suzuki et al. 2015), therefore transcripts with adenine and guanine at the wobble position were screened for the threefold to sixfold-degenerate amino acids in the available transcriptome reads (Table S6). *S. rosetta* appears to lack a $tRNA^{Thr}_{AGT}$ gene, therefore only tRNA molecules for leucine, isoleucine, valine, serine, proline, alanine and arginine were screened for in this species. The number of tRNA transcripts identified in the transcriptomes were low, however across the two species tRNA molecules with guanine at the wobble position were identified for isoleucine, serine, proline, alanine and arginine (Table S6). tRNA genes with guanine at the wobble position for those amino acids are absent from the genomes, therefore it can be concluded that the holozoan protists are modifying their tRNA molecules through adenosine deamination. If, as proposed here, adenosine in tRNA anticodons is deaminated to inosine for all high degeneracy amino acids, the major tRNA gene for each amino acid will be complementary to the most frequently used optimal codons in

all cases with the exception of threonine in the two choanoflagellates and lysine in *M. brevicollis*.

One unexpected finding was the identification of a putative tRNA gene for the amino acid selenocysteine in *C. owczarzaki* (Table S4), which possesses the anticodon TCA that should complement UGA codons. Analysis by tRNAScan-SE indicated that the $tRNA^{Sec}$ gene is not a pseudogene, however the screening of both choanoflagellate genomes failed to identify any $tRNA^{Sec}$ genes.

Codon Usage Bias Increases with Expression Level in *S. rosetta* and *C. owczarzaki*

Deep coverage transcriptome datasets are available for both *S. rosetta* and *C. owczarzaki*, although not for *M. brevicollis*, and the number of reads for each gene was plotted against F_{op} for each gene. Figure 4 shows the relationships when genes with very low expression levels (less than 100 reads, which is similar to the level of pseudogene transcription observed in *C. owczarzaki* by Carr and Suga (2014)) have been excluded. In both species, there is a general trend observed for F_{op} to increase as the level of expression increases. The data therefore are consistent with selection operating upon translational efficiency, with highly expressed genes preferentially using optimal codons in comparison to weakly expressed genes. However, in both species R^2 is low (0.218 for *S. rosetta* and 0.231 for *C. owczarzaki* after excluding genes with less than 100 reads), indicating that evolutionary forces in addition to selection for translational efficiency are operating in the genomes of *S. rosetta* and *C. owczarzaki*.

Protein Domain Codons Exhibit Stronger Codon Usage Bias Than Non-Domain Codons

In order to determine if translational accuracy also plays a role, codon usage bias was examined in gene regions that encode structural or functional domains, as well as non-domain codons. Domains for each gene in the three bias categories were identified using the NCBI gene annotation and the degree of codon usage bias was determined using F_{op}.

F_{op} was significantly elevated in domain codons compared to non-domain codons in all three categories (*t* test, $P<0.0001$ for each comparison) for all three species (Figure 5, Table S7). The data therefore are consistent with selection for translational accuracy being a driver of codon usage in the three holozoan protists and that its effects are observable even in weakly biased genes.

Estimates of the Strength of Translational Selection in the Three Holozoan Protists

All three genomes show evidence for selection acting upon translational accuracy and both *S. rosetta* and *C. owczarzaki*, for which expression data are available, show evidence for selection for translational efficiency. Numerous methodologies have been devised to determine the strength of selection acting upon codon usage and two are used here in an attempt to quantify selection within the holozoan protists. Eyre-Walker and Bulmer (1995) used the odds ratio for pairs of synonymous codons in high and low expressed genes (Equation 4 in that paper) to measure the strength of selection. This is adapted here to compare high and low bias genes, based upon Axis 1 of the correspondence analyses, since expression data are only available for two of the three species. Direct comparisons must be taken with caution, due to the different methodologies used to generate odd ratios, however the values produced for twofold

357 degenerate amino acids (Table S8) are similar for the holozoan protists to those for
358 enterobacteria determined by Eyre-Walker and Bulmer.

359 Sharp et al. (2005) developed a population genetics based model that results in the
360 statistic *S*, which estimates the strength of selected codon usage bias. This was used to
361 study selection in a broad range of bacterial taxa, whilst dos Reis and Wernisch
362 (2009) noted that *S* is the log of Eyre-Walker and Bulmer's odd ratio when estimating
363 the strength of selection on codon usage in eukaryotes. This calculation is applied
364 here to estimate *S* for each twofold amino acid, as well as for the genome using the
365 weighted average for each twofold degenerate amino acid in the highly biased genes
366 (Tables 1 and S8). The genome averages ($\hat{S}$) for *S. rosetta* and *C. owczarzaki* are
367 similar, if a little lower, to the estimates of dos Reis and Wernisch (2009) for Dikarya
368 fungi. *M. brevicollis* exhibits a lower value for $\hat{S}$, however it is comparable to values
369 from ecdysozoans (dos Reis and Wernisch (2009), which have previously been shown
370 to have selection upon their codon usage (reviewed in Duret 2000).

371 **Discussion**

372 **Codon Usage is Highly Conserved Across the Three Holozoan Protists**

373 Codon usage has been extensively studied in a broad range of opisthokont
374 taxa, mainly centring on metazoans and fungi. The sequencing of whole genomes
375 from unicellular holozoans provides an insight into the evolutionary forces that drive
376 codon usage in the closest relatives of Metazoa. Molecular clock estimates suggest
377 that choanoflagellates last shared a common ancestor with *C. owczarzaki* over one
378 billion years ago (Parfrey et al. 2011). Despite this antiquity, many of the aspects of
379 codon usage across the three holozoan protists examined are remarkably similar.

All three protists studied show a similar bias towards GC-ending codons and optimal codons which mainly end in either cytosine or guanine. Within fourfold and sixfold-degenerate amino acids, there is a clear preference for cytosine at the synonymous third position of codons. Every amino acid shares at least one optimal codon across the three species and only seven amino acids show any variation in their full complement of optimal codons (Table 2). This suggests that codon usage is either highly conserved, or that there has been an extraordinary level of convergent evolution between the species.

One area in which *C. owczarzaki* differs from the two choanoflagellates is the presence of a tRNA gene for the amino acid selenocysteine (Table S4). The presence in the *C. owczarzaki* genome of $tRNA^{Sec}$, which possesses an anticodon that complements UGA stop codons, suggests that the current annotation of selenoproteins may require revision; selenocysteine codons may have been identified as stop codons, resulting in erroneously truncated conceptual proteins. The 3' untranslated regions of selenoprotein genes contain a Sec incorporation sequence (SECIS), which is required for the correct incorporation of selenocysteine in place of translation termination (reviewed in Shetty et al. 2014). SECIS screening the putative selenoprotein genes in the *C. owczarzaki* genome could highlight any proteins which have been misannotated with premature stop codons.

A second potential difference between species is the approximately 200 genes in the *M. brevicollis* genome which do not conform to the general trend of increasing GC3s with increasing bias (Figures 2 and S1). Further investigation of the genes, in particular undertaking reciprocal BLAST analyses, indicated that many of the genes may actually be artefacts generated during genome annotation. The use of *Nc* plots

has the potential to aid genome annotation, particularly in species which exhibit strong trends in codon usage bias as observed here. Genes which are candidates for being false positives in genome annotation can be scrutinized in greater depth in order to determine if they are genuine. Furthermore, *Nc* plots may also have a role in identifying genes that are present due to horizontal transfer. Choanoflagellates are known to have acquired a large number of genes from algae and bacteria (Tucker 2013), possibly through the escape of prey DNA from food vacuoles into their nuclei. In cases where donor species have distinct codon usage from the recipient species, horizontally transferred genes are likely to be placed away from vertically inherited genes on *Nc* plots. This methodology however would be limited in only identifying recent transfers, as successfully transferred genes are likely to adapt their codon usage to that of their new host.

As is set out below, the forces that determine codon usage within the three holozoan protists appear to be extremely similar. Since the last common ancestor of choanoflagellates and filastereans was also a direct ancestor of Metazoa, the analyses presented here show how codon usage was likely to have been driven in unicellular premetazoans.

Selection for Optimal Codons is a Major Driver of Codon Usage Bias

There was a significant reduction in GC3s when comparing mid-bias to high-bias genes and low-bias to mid-bias genes, consistent with the enriched use of translationally optimal, GC-ending codons in highly biased genes. The non-coding GC content of the genes did not show the same pattern of significant reductions across gene bias categories, indicating that mutation pressure is not a major driver of codon usage bias in the genes of the three species.

428 The stop codons employed by highly biased genes do not show evidence of
429 GC mutation pressure. Whilst the sense codons of highly biased genes show a strong
430 for preference for GC at synonymous sites, the preferred stop codon of all three
431 species is UAA. Brown et al (1990) noted that highly expressed genes in a broad
432 variety of eukaryotes prefer UAA as a stop codon, however they also noted that
433 species with GC-rich genomes had a preference for UGA as a stop codon. The only
434 exception to GC-rich species preferring UGA was the unicellular *Chlamydomonas*
435 *reinhardtii*, which exhibits a similar pattern to the three holzoan protists here in
436 having a GC-rich genome and a preference for UAA stop codons in highly expressed
437 genes. Brown et al. (1990) proposed a possible selective advantage in using UAA as a
438 stop codon, since UGA may be misread as a UGG tryptophan codon. Furthermore,
439 UGA and UAG are more likely than UAA to act as suppressible stop codons (Brown
440 et al. 1990). Similar selective forces may be in operation within the GC-biased
441 holozoan protist genomes, providing an explanation for the excess of UAA stop
442 codons in highly biased genes.

443 Gene expression data from *S. rosetta* and *C. owczarzaki* were consistent with
444 selection for translational efficiency in both species. Expression data are currently
445 unavailable for *M. brevicollis*, but the KOG category that includes highly expressed
446 protein translation genes is significantly enriched in high bias genes within this
447 species (Table S1). Furthermore, across all three species, the major tRNA genes
448 appear to be complementary, albeit after tRNA modification, with optimal codons.
449 This also points to selection for translational efficiency, as it appears likely that the
450 most abundant tRNA molecules in each species will be available to bind to optimal
451 codons. The codon usage of highly biased genes and the number of tRNA genes

452 therefore appear to have co-evolved for the rapid synthesis of proteins within
453 holozoan protists.

454 Genes with low expression levels show lower levels of enrichment for optimal
455 codons, consistent with the predictions of selection for translational efficiency.
456 However, even genes that show weak codon usage bias exhibit the signature of
457 selection for translational accuracy, due to regions encoding functional domains being
458 enriched for optimal codons, which is in agreement with recent findings in weakly
459 expressed *E. coli* genes (Yannai et al. 2018). Selection for translational accuracy
460 therefore also appears to be a driver of codon usage bias in choanoflagellates and
461 filastereans.

462 Despite the different life histories of the free-living choanoflagellates and the
463 symbiotic *C. owczarzaki*, all three holozoans exhibit similar patterns within their
464 codon usage bias. If *C. owczarzaki* does undergo population bottlenecks when it
465 passes from one host to another, they are not sufficiently severe to significantly
466 dampen selection for optimal codons. Indeed, the estimated strength of selected codon
467 usage bias ($\hat{S}$), based upon twofold degenerate amino acids, is higher for *C.*
468 *owczarzaki* than for either choanoflagellate. A further potential source of difference
469 between the choanoflagellates and *C. owczarzaki* is the use of EFL as an elongation
470 factor in the former, whilst the latter employs EF1A (Carr et al. 2017). The elongation
471 factors facilitate the delivery of aminoacyl-tRNA molecules to the ribosome (Riis et
472 al. 1990) and it could be speculated that the two proteins may exhibit differences in
473 how they interact with tRNA molecules. However, as with the differences in life
474 histories, the alternative elongation factors do not appear to have a noticeable impact
475 on the tRNA genes or optimal codons of the holozoan protists.

476 Within the two choanoflagellates codon usage bias is stronger in *S. rosetta*
477 than *M. brevicollis*. The presence of possible inaccurately annotated genes in *M.*
478 *brevicollis* may however be acting to lower the apparent strength of codon usage bias
479 within this species. If selection is genuinely weaker in *M. brevicollis*, the stronger bias
480 in *S. rosetta* indicates that it may have a larger effective, if not absolute, population
481 size. There is currently a lack of empirical evidence to test this hypothesis, with no
482 ecological or population nucleotide diversity data available for either choanoflagellate
483 species. Recent studies have shown that *S. rosetta* undergoes sexual reproduction
484 (Levin and King 2013; Woznica et al. 2017), but, whilst there is indirect evidence of
485 sexual reproduction in both *M. brevicollis* and *C. owczarzaki* (Carr et al. 2010; Suga
486 et al. 2014), it has yet to be confirmed in either species. Differences in recombination
487 rates, as a result of differences in the frequency of sexual reproduction, may influence
488 the effective population sizes of the three holozoan protists and therefore the
489 efficiency of selection upon optimal codons. The rates of cell division between the
490 two species of choanoflagellate are also currently unknown and this component of life
491 history may have an impact upon the strength of selection of translational efficiency,
492 since rapid cell division is likely to require rapid protein synthesis. *M. brevicollis* is a
493 strictly unicellular choanoflagellate (reviewed in Carr et al. 2017) and cellular
494 differentiation has to be reported in this species; in contrast *S. rosetta* can form
495 ephemeral colonies and produce at least five different cell types (Dayel et al. 2011).
496 Whether the different *S. rosetta* cell types exhibit similar rates of cell division, how
497 the rates compare to those of *M. brevicollis* and whether this influences selection on
498 codon usage highlights the lack of current knowledge on much of choanoflagellate
499 biology.

Evidence for Widespread Deamination of Adenosine at the Wobble Position of tRNA Molecules

The major tRNA genes of twofold degenerate amino acids show a perfect complementary match to their optimal codons with the exception of lysine in *M. brevicollis*. Analysis of the *M. brevicollis* genome with tRNA-Scan-SE identified two $tRNA^{Lys}$ genes, with one gene complementing each lysine codon (Table S4). In addition, the screen of the genome also identified five putative $tRNA^{Lys}$ pseudogenes, all of which had CUU anticodons. The lack of a match between optimal codon and major tRNA for lysine may therefore have only arisen recently with the loss of function in the $tRNA^{Lys}_{CUU}$ pseudogenes.

Of the amino acids with higher levels of degeneracy in their genetic codes, tRNA genes with adenosine at the wobble position make up almost all of the major tRNA genes (Tables 2 and S3). The most frequently used optimal codons for these amino acids show cytosine at the degenerate position, with the exception of UCG, which is the most frequently used optimal codon for serine in *C. owczarzaki*. It is clear that these optimal codons do not rely upon standard Watson-Crick base pairings to bind tRNA, since tRNA genes with guanine at the wobble position of anticodons are absent (Table S5). Complementation of codons and anticodons could be achieved through tRNA modification, with adenosine at the wobble position of the tRNA being converted to inosine. Rafels-Ybern et al. (2017) recently showed that the deamination of this adenosine is widespread amongst eukaryotes and that codons complementary to deaminated tRNA molecules are enriched in highly expressed genes. Glycine is the only amino acid where the phenomenon does not occur in eukaryotes, due to a lack of stability in the structure of $tRNA^{Gly}_{ACC}$ molecules (Yokoyama and Nishimura 1995;

524 Saint-Léger et al. 2016). Glycine is also the only higher degeneracy amino acid which
525 shows standard Watson-Crick base pairing between major tRNA gene and the most
526 frequent optimal codon in the studied genomes (Table 2). Transcriptome sequences
527 contain cDNA sequences with guanine at the wobble position of anticodons. It
528 therefore appears that adenosine modification is the most plausible explanation for the
529 presence of these tRNA molecules in the transcriptomes. The transcriptome sequences
530 also show the presence of tRNA molecules with adenine at the wobble position of
531 anticodons, which have escaped deamination (Table S6). Those tRNAs will bind to
532 uracil-ending codons under standard Watson-Crick base pairing. This is consistent
533 with the occurrence of lower frequency, uracil-ending optimal codons for some higher
534 degeneracy amino acids (Table S3) and provides a selection-based explanation for
535 uracil-ending optimal codons.

536 The wide-scale deamination of adenosine to inosine is common across
537 different RNA types in Metazoa (Porath et al. 2017) and Rafels-Ybern et al. (2017)
538 showed that metazoan and plant genomes possess a greater concentration of codons
539 translated by modified tRNAs than the genomes of unicellular eukaryotes. From this
540 observation, they speculated that such codons were important in the evolution of
541 multicellularity in metazoans and plants. However, based upon the analyses presented
542 here, it appears that the large-scale usage of deaminated, tRNAs molecules evolved
543 early within Holozoa prior to the emergence of true multicellularity.

544 **Optimizing Future Transgene Design Within Holozoan Protists**

545 The three holozoan protists studied here are lab workhorses in the study of the
546 evolution of Holozoa and the origin of metazoan multicellularity (Abedin and King
547 2008; Carr et al. 2010; Najile et al. 2016; Sebé-Pedrós et al. 2016; Woznica et al.

548 2016). Transgenics systems are currently being developed for holozoan protists
549 (Parra-Acero et al. 2018), and the data presented here provide important information
550 for the expression of artificial transgenes. The use of a species' optimal codons in a
551 transgene can increase the yield of the encoded protein by three orders of magnitude
552 (Gustafsson et al. 2012) and Table 2 presents the optimal codons, as well as preferred
553 stop codons, for all three species, enabling the future tailored design of transgenes for
554 each of the holozoans.

555 **Conclusions**

556 The genome analyses presented highlight the role of natural selection in the
557 evolution of codon usage bias within extant holozoan protists. Importantly, the data
558 also contribute to the growing knowledge of unicellular, premetazoan evolution. The
559 data indicate that natural selection operating upon both translational accuracy and
560 efficiency resulted in a GC-bias in the codon usage bias of premetazoans. It is
561 increasingly clear that the modification of RNA molecules through adenosine
562 deamination to inosine is important in extant metazoans. However, it can be seen that
563 tRNA molecules for higher level degeneracy amino acids underwent deamination in
564 the very earliest premetazoans and that this trait evolved prior to the divergence of
565 Filasterea from the lineage that lead to both choanoflagellates and metazoans.
566 Furthermore, highly biased genes in early metazoan evolutionary history would have
567 been enriched with codons that complemented the modified tRNA molecules.

568 The transition from unicellular premetazoan to multicellular stem-group metazoan is
569 likely to have resulted in a number of changes that led to a reduction of efficient
570 selection for optimal codons. In particular, the effective population sizes of metazoans
571 are believed to be lower than those of unicellular species, resulting in less efficient

572 selection. Longer generation times, due to the division of somatic and germ line cells
573 and the development of adult life-stages, are also likely to have reduced selection for
574 rapid cell division, thereby relaxing selective pressures for translational efficiency.
575 Finally, a differentiated multicellular bodyplan allowed the evolution of tissue-
576 specific tRNA expression patterns, breaking the link between tRNA gene number and
577 expression level.

578 **Methods & Materials**

579 All analysed datasets are available from the corresponding author upon
580 request.

581 **Codon Usage Statistics**. Complete annotated transcriptome sequences for *M*.
582 *brevicollis* and *S*. *rosetta* were downloaded from the Origins of Multicellularity
583 Project at the Broad Institute; the *C*. *owczarzaki* transcriptome was downloaded from
584 the EnsemblProtists database. The versions used were:
585 monosiga_brevicollis_mx1_1_transcripts, salpingoeca_rosetta_1_transcripts and
586 Capsaspora_owczarzaki_atcc_30864.C_owczarzaki_V2.cds. Codon usage statistics
587 were determined in CodonW (Peden 1999). Optimal codon, fop.coa files, were
588 generated by correspondence analyses, using RSCU, of the complete transcriptomes
589 of each species using default parameters. *S*. *rosetta* and *C*. *owczarzaki* both have large
590 publicly available datasets of gene expression (see Gene Expression in *S*. *rosetta* and
591 *C*. *owczarzaki* below) and in both species the principal axis of the correspondence
592 analysis shows the strongest relationship with expression, consistent with selection
593 operating upon translational efficiency (Peden, 1999). Optimal codons were also
594 determined using expression levels (Lloyd and Sharp 1991, 1993; Stenico et al. 1994),
595 by identifying codons present at significantly higher frequencies in the 5% most

highly expressed genes compared to the 5% most weakly expressed genes. Values of *Nc*, GC3s and F_{op} were calculated in CodonW. F_{op} values were determined using the fop.coa file produced by CodonW for *M. brevicollis* and the optimal codons determined using expression levels for *S. rosetta* and *C. owczarzaki*.

Odds ratios between synonymous codon pairs within the 5% most biased genes and the least biased genes, based upon the primary axes of the CodonW correspondence analyses were determined for all twofold degenerate amino acids. The population parameter for the strength of codon usage bias, *S*, was estimated for each of the nine twofold degenerate amino acid by determining the log of the odds ratio values. Species estimates, $\hat{S}$, were calculated by determining the weighted average for amino acids based upon the number of codons in highly biased genes (based upon Sharp et al. 2005).

Determining Non-Coding GC Content

The genes in each bias-category were also screened for predicted introns in the NCBI gene annotations. When possible 200bp of both 5' and 3' flanking DNA was extracted for each in the three bias categories. Where predicted intergenic regions were less than 200bp, or genes were located at the ends of genomic contigs, the maximum possible length of flanking DNA was extracted. All introns and flanking DNA for each gene were extracted, concatenated and their total GC content was calculated in CodonW. Mean and standard deviation values were then generated for each category.

tRNA Gene Screening.

618 The scaffolds of each annotated genome were downloaded from NCBI. The
619 *M. brevicollis* dataset was made up from the 218 scaffolds of the version 1 genome
620 assembly, *S. rosetta* dataset comprised the 3,086 scaffolds of the version 1 assembly
621 and the *C. owczarzaki* dataset was the 84 scaffolds from the version 2 assembly. The
622 program tRNAscan-SE 2.0 (Lowe and Eddy, 1997), applied through an online server
623 (Lowe and Chan 2016), was used to identify tRNA genes using default settings.

624 **Gene Expression in *S. rosetta* and *C. owczarzaki***

625 The level of gene expression for each species was determined by the
626 examining the high, medium and low bias categories, using SMALT v. 0.2.6 (Hannes
627 Ponstingl, Genome Research Ltd) in both *S. rosetta* and *C. owczarzaki*. For *S. rosetta*
628 the transcriptomic SRA files SRX042046-SRX042054 (122.1 million reads) and for
629 *C. owczarzaki* SRX1690425-SRX1690428 and SRX155789-SRX155797 (665.2
630 million reads) were downloaded from the NCBI SRA database. SRA reads were
631 mapped onto each gene sequence to determine expression level. The number of reads
632 for each gene was then calculated in Tablet v. 1.16.09.06 (Milne et al. 2013). Reads
633 were additionalyy mapped on to the tRNA genes of alanine, arginine, isoleucine,
634 leucine, proline, serine, threonine and valine that possess adenosine and guanosine at
635 the wobble position.

636 **Optimal Codon Usage in Domain and Non-Domain Codons**.

637 Within the bias-categories, each gene was divided into putative functional
638 domain and non-domain codons. Codons encoding functional domains were identified
639 from the annotated Regions within the NCBI file for each gene. Genes which did not
640 encode annotated regions were excluded from the analyses; furthermore, genes were
641 excluded when the defined region spanned the entire gene, or all codons with the

642 exception of the start and/or the stop codon. F_{op} values were determined in CodonW
643 for the functional domain and non-domain regions. Values of *Nc* were not considered,
644 as some regions did not contain codons from all of the required redundancy
645 categories.

646 **Acknowledgements**

647 JS is supported by a Leverhulme Trust Doctoral Scholarship. The choanoflagellate
648 images used in Figure 1 were kindly provided by Tom Davidson. We are thankful to
649 three anonymous reviewers for comments upon our manuscript.

650 **References**

651 Abedin M, King N. 2008. The Premetazoan Ancestry of Cadherins. Science 319: 946-
652 948.

653 Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural
654 selection and translational accuracy. Genetics 136: 927-935.

655 Akashi H. 1995. Inferring weak selection from patterns of polymorphism and
656 divergence at silent sites in *Drosophila* DNA. Genetics 139: 1067-1076.

657 Brown CM, Stockwell PA, Trotman CNA, Tate WP. 1990. Sequence analysis
658 suggests that tetra-nucleotides signal the termination of protein synthesis in
659 eukaryotes. Nucleic Acids Res 18: 6339-6345

660 Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage.
661 Genetics 129: 897-907.

662 Carr M, Leadbeater BSC, Hassan R, Nelson M, Baldauf SL. 2008. Molecular
663 phylogeny of choanoflagellates, the sister group to Metazoa. Proc Natl Acad Sci USA
664 105: 16641-16646.

665 Carr M, Leadbeater BSC and Baldauf SL. 2010. Conserved meiotic genes point to sex
666 in the choanoflagellates. J Eukaryot Microbiol 57: 56-62.

667 Carr M, Richter DJ, Fozouni P, Smith TJ, Jeuck A, Leadbeater BSC, Nitsche F. 2017.
668 A six-gene phylogeny provides new insights into choanoflagellate evolution. Mol
669 Phylogenet Evol 107: 166-178.

670 Carr M, Suga H. The holozoan *Capsaspora owczarzaki* possesses a diverse
671 complement of active transposable element families. Genome Biol Evol 6: 949-963.

672 Clarke B. 1970. Darwinian evolution of proteins. Science 168: 1009-1011.

673 Dayel MJ, Alegado RA, Fairclough SR, Levin TC, Nichols SA, McDonald K, King
674 N. 2011. Cell differentiation and morphogenesis in the colony-forming
675 choanoflagellate *Salpingoeca rosetta*. Dev Biol 357: 73-82.

676 dos Reis M, Wernisch L. 2009. Estimating translational selection in eukaryotic
677 genomes. Mol Biol Evol 26: 451-461.

678 Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-
679 adapted for optimal translation of highly expressed genes. Trends Genet 16: 287-289.

680 Ehrenberg M, Kurland CG. 1984. Costs of accuracy determined by a maximal growth
681 rate constraint. Q Rev Biophys 17: 45-82.

682 Eyre-Walker A, Bulmer M. 1995. Synonymous substitution rates in enterobacteria.
683 Genetics 140: 1407-1412.

684 Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E,
685 Richter DJ, Russ C, Westbrook MJ et al. 2013. Premetazoan genome evolution and
686 the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*.
687 Genome Biol 14: R15.

688 Figuet E, Ballenghien M, Romiguier J, Galtier N. 2014. Biased gene conversion and
689 GC-content evolution in the coding squences of reptiles and vertebrates. Genome Biol
690 Evol 7:240-250.

691 Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene
692 expressivity. Nucleic Acids Res 10: 7055-7074.

693 Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and
694 the genome hypothesis. Nucleic Acid Res 8: r49-r62.

695 Gustafsson C, Minshull J, Govindarajan S, Ness J, Villalobos A, Welch M. 2012.
696 Engineering genes for predictable protein expression. Protein Expr Purif 83: 37-46.

697 Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias.
698 Genetics 138: 227-23.

699 Hehenberger E, Tikhonenkov DV, Kolisko M, Del Campo J, Esaulov AS, Mylnikov
700 AP, Keeling PJ. 2017. Novel predators reshape holozoan phylogeny and reveal the
701 presence of a two-component signalling system in the ancestor of animals. Curr Biol
702 27: 2043-2050.

703 Hertel LA, Bayne CJ, Loker ES. 2002. The symbiont *Capsaspora owczarzaki*, nov.
704 gen. nov. sp., isolated from three strains of the pulmonate snail *Biomphalaria*
705 *glabrata* is related to members of the Mesomycetozoea. Int J Parasitol 32: 1183-1191.

706 Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer
707 RNAs and the occurrence of the respective codons in its protein genes: A proposal for
708 a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol
709 Biol 151: 389-409.

710 King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S,
711 Hellsten U, Isogai Y, Letunic I et al. 2008. The genome of the choanoflagellate
712 *Monosiga brevicollis* and the origin of metazoans. Nature 451:783-788.

713 Kliman RM, Hey J. 1994. The effects of mutation and natural selection on codon bias
714 in the genes of Drosophila. Genetics 137: 1049-1056.

715 Leadbeater BSC. 2015. *The Choanoflagellates: Evolution, Biology and Ecology*.
716 Cambridge University Press.

717 Levin TC, King N. Evidence for sex and recombination in the choanoflagellate
718 *Salpingoeca rosetta*. Current Biology 23: 1-5.

719 Liu G, Tang J, Edmonds BT, Murray J, Levin S, Condeelis J. 1996. F-actin sequesters
720 elongation factor 1a from interaction with aminoacyl-tRNA in a pH- dependent
721 reaction. J Cell Biol 135:953–963.

722 Lloyd AT, Sharp PM. 1991. Codon usage in *Aspergillus nidulans*. Molec Gen Genet
723 230: 288-294.

724 Lloyd AT, Sharp PM. 1993. Synonymous codon usage in *Kluyveromyces lactis*. Yeast
725 9: 1219-1228.

726 Lowe TM, Chan, PP. 2016. tRNAscan-SE On-line: integrating search and context for
727 analysis of transfer RNA genes. Nucleic Acids Res 44: W54-W57.

728 Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of
729 transfer RNA genes in genomic sequence. Nucl Acids Res 25: 955-964.

730 Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302: 1401-
731 1404.

732 Maside X, Lee AW, Charlesworth B. 2004. Selection on codon usage in *Drosophila*
733 *americana*. Curr Biol 14: 150-154.

734 Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D.
735 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief
736 Bioinform 14: 193-202.

737 Najile SR, Molina MC, Ruiz-Trillo I, Uttaro AD. 2016. Sterol metabolism in the
738 filasterean *Capsaspora owczarzaki* has features that resemble both fungi and animals.
739 Open Biol 6: 160029.

740 Nitsche F, Carr M, Arndt H, Leadbeater BSC. 2011. Higher level taxonomy and
741 molecular phylogenetics of the Choanoflagellatea. Journal of Eukaryotic
742 Microbiology 58: 452-462.

743 Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early
744 eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA
745 108: 13624-13629.

746 Parra-Acero H, Ros-Rocher N, Perez-Posada A, Kożyczkowska A, Sánchez-Pons N,
747 Nakata A, Suga H, Najle SR, Ruiz-Trillo I. Transfection of *Capsaspora owczarzaki*, a
748 close unicellular relative of animals. Development 145: dev162107.

749 Patterson DJ, Nygaard K, Steinberg G, Turley CM. 1993. Heterotrophic flagellates
750 and other protists associated with oceanic detritus throughout the water column in the
751 mid North Atlantic. J Mar Biol Ass UK 73: 67-95.

752 Peden JF. 1999. Analysis of codon usage. PhD Thesis, University of Nottingham

753 Pedersen S. 1984. *Escherichia coli* ribosome's translate *in vivo* with variable rate.
754 EMBO J 3: 2895-2898.

755 Porath HT, Knisbacher BA, Eisenberg E, Levanon EY. 2017. Massive A-to-I RNA
756 editing is common across the Metazoa and correlates with dsRNA abundance.
757 Genome Biol 18: 185.

758 Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of
759 codon identity and context. J Biol Chem 262: 11351-11356.

760 Rafels-Ybern A, Torres AG, Grau-Bove X, Ruiz-Trillo I, Ribas de Pouplana L. 2017.
761 Codon adaptation to tRNAs with inosine modification at position 34 is widespread
762 among Eukaryotes and present in two Bacterial phyla. Online Early, RNA Biology

763 Riis B, Rattan SI, Clark BF, Merrick WC. 1990. Eukaryotic protein elongation
764 factors. Trends Biochem Sci 15: 420-424.

765 Saint-Léger A, Bello C, Dans PD, Torres AG, Novoa EM, Camacho N, Orozco M,
766 Kondrashov FA, de Pouplana LR 2016. Saturation of recognition elements blocks
767 evolution of new tRNA identities. Sci Adv 2: e1501860.

768 Sebé-Pedrós A, Peña MI, Capella-Gutiérrez S, Antó M, Gabaldón T, Ruiz-Trillo I,
769 Sabidó E. 2016. High-throughput proteomics reveals the unicellular roots of animal
770 phosphosignalling and cell differentiation. Dev Cell 39: 186-197.

771 Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS,
772 Cavalier-Smith T. 2008. Multigene phylogeny of the Choanozoa and the origin of
773 animals. PLoS ONE 3: e2098.

774 Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA sequence
775 evolution: the sounds of silence. Phil Trans R Soc Lond B 349:241-247.

776 Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the
777 strength of selected codon usage bias among bacteria. Nucleic Acids Res 33: 1141-
778 1153.

779 Shetty SP, Shah R, Copeland PR. 2014. Regulation of selenocysteine incorporation
780 into the selenium transport protein Selenoprotein P*. J Biol Chem 289: 25317-25326.

781 Smith NGC, Eyre-Walker A. 2001. Synonymous codon bias ins not caused by
782 mutation bias in G+C-rich genes in humans. Mol Biol Evol 18:982-986.

783 Snoke MS, Berendonk TU, Barth D, Lynch M. 2006. Large global effective
784 population sizes in *Paramecium*. Mol Biol Evol 23: 2474-2479.

785 Sørensen MA, Pedersen S. 1991. Absolute in vivo translation rates of individual
786 codons in *Escherichia coli*: the the two glutamic acids codons GAA and GAG are
787 translated with a threefold difference in rate. J Mol Biol 222: 265-280.

788 Stenico M, Lloyd AT, Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*:
789 delineation of translational selection and mutational biases. Nucleic Acids Res 22:
790 2437-2446.

791 Stibbs HH, Owczarzak A, Bayne CJ, DeWan P. 1979. Schistosome sporocyst-killing
792 amoebae isolated from *Biomphalaria glabrata*. J Invertebr Pathol 33: 159-170.

793 Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M,
794 Kerner P, Vervoort M, Sánchez-Pons N. 2013. The *Capsaspora* genome reveals a
795 complex unicellular prehistory of animals. Nat Commun 4:2325.

796 Suzuki T, Ueda H, Okada S, Sakurai M. 2015. Transcriptome-wide identification of
797 adenosine-to-inosine editing using the ICE-seq method. Nat Protoc 10: 715-732.

798 Tong SM. 1997. Heterotrophic flagellates and other protists from Southampton
799 Water, UK. Ophelia 47: 71-131.

800 Tucker RP. 2013. Horizontal gene transfer in choanoflagellates. J Exp Zool (Mol Dev
801 Evol) 9999B: 1-9.

802 Watts PC, Lundholm N, Ribeiro S, Ellegaard M. 2013. A century-long genetic record
803 reveals that protist effective population sizes are comparable to those of macroscopic
804 species. Biol Lett 9: 20130849.

805 Woznica A, Cantley AM, Beemelmanns C, Freinkman E, Clardy J, King N. 2016.
806 Bacterial lipids activate, synergize, and inhibit a developmental switch in
807 choanoflagellates. Proc Natl Acad Sci USA 113: 7894-7899.

808 Woznica A, Gerdt JP, Hulett RE, Clardy J, King N. 2017. Mating in the closest living
809 relatives of animals is induced by a bacterial chondroitinase. Cell 170: 1175-1183.

810 Wright F. 1990. The 'effective number of codons' used in a gene. Gene 87: 23-29.

811 Yannai A, Katz S, Hershberg R. 2018. The codon usage of lowly expressed genes in
812 subject to natural selection. Genome Biol Evol 10: 1237-1246.

813 Yokoyama S, Nishimura S. 1995. Modified nucleosides and codon recognition in
814 tRNA: Structure, Biosynthesis and Function. Eds Söll D & RajBhandry U. American
815 Society for Microbiology, Washington, DC.

816

817

818 **Figure Legends**

819 **Fig. 1. Simplified phylogeny of Holozoa.** Lineages which show conserved ancestral
820 traits are shown on blue branches. The red branch represents the loss of ancestral
821 traits and the dotted blue/red line shows transitional stages during the process of loss.
822 Approximate divergence dates are taken from Parfrey et al. (2011).

823

824 **Fig. 2. *Nc* plots for *M. brevicollis*, *S. rosetta* and *C. owczarzaki*.** GC3s values are
825 shown on the x-axis and *Nc* values are given on the y-axis. The curved line on each
826 plot represents the expected position of genes evolving under a neutral mutation
827 model (Wright 1990).

828

829 **Fig. 3. Comparisons of mean GC3s and non-coding DNA GC-content for the**
830 **holozoan protists.** The bars represent the standard deviation of the values from each
831 category. The GC3s values are on the left y-axis (purple dot: highly biased; dark blue:
832 mid bias; light blue: low bias). Non-coding GC content is shown on the right y-axis
833 axis (Flanking DNA. Dark green dot: highly biased; light green: mid bias; mustard:
834 low bias. Intron DNA. Red dot: highly biased; orange: mid bias; yellow: low bias).

835

836 **Fig. 4. Relationships between gene expression levels and F_{op} in *S. rosetta* and *C.**
837 ***owczarzaki*.** The x-axis (number of reads per gene) and line of best fit are both shown
838 with a logarithmic scale for both species.

839

Fig. 5. Mean F_{op} values in domain encoding and non-domain encoding codons. The dark blue dot gives the mean value for domain codons and the red dot shows the mean value for non-domain codons in highly-biased genes. The light blue dot gives the mean value for domain codons and the orange dot shows the mean value for non-domain codons in mid-biased genes. The green dot gives the mean value for domain codons and the yellow dot shows the mean value for non-domain codons in low - biased genes. The bars represent the standard deviation of the values from each category. **** - $P<0.0001$.

Table 1. Whole genome and codon usage statistics in the transcriptomes of the three holozoan protists.

Species	Genome Size (Mb)[a]	Number of CDS Sequences	CDS % Genome[a]	GC-Content	GC3s (±sd)	*Nc* (±sd)	F_{op} (±sd)	$\hat{S}$
M. brevicollis	41.6	9,171	39.7	0.549	0.638±0.060	48.05±5.62	0.572±0.080	1.27
S. rosetta	55.0	11,736	43.5	0.566	0.707±0.073	44.78±5.36	0.576±0.079	2.00
C. owczarzaki	28.0	10,123	58.7	0.538	0.653±0.075	47.60±6.45	0.494±0.100	2.18

a. Data taken from Fairclough et al. (2013) and Suga et al. (2014)

Note. Abbreviations. Mb: Megabases; CDS: Coding Sequence; GC3s: guanine+cytosine content at synonymous third positions; *Nc*: the effective number of codons; F_{op}: the frequency of optimal codons; $\hat{S}$: the strength of selected codon usage bias.

855 **Table 2.** Optimal codons designated for the three species of holozoan protist.

Amino Acid	***M. brevicollis***		***S. rosetta***		***C. owczarzaki***	
	Expression-Level	**CodonW**	**Expression-Level**	**CodonW**	**Expression-Level**	**CodonW**
Phe	n/a	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**
Leu	n/a	CUC*, CUG	CUC*, CUG	CUC*, CUG	CUC*	CUC*
Ile	n/a	**AUU**, AUC*	**AUU**	AUC*	AUC*	AUC*
Met	n/a	AUG	AUG	AUG	AUG	AUG
Val	n/a	GUC*	GUC*	GUC*	GUC*	GUC*
Ser	n/a	AGC, UCC*, UCG	**UCU**, UCC*, UCG	AGC, **UCU**, UCC*, UCG	UCC*, UCG	**UCU**, UCC*, UCG
Pro	n/a	CCC*	CCC*, CCG	CCC*, CCG	CCC*	CCC*
Thr	n/a	ACC	ACC, ACG	ACC, ACG	ACC*	ACC*
Ala	n/a	GCC*	GCC*	GCC*	GCC*	GCC*

Tyr	n/a	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**
His	n/a	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**
Gln	n/a	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**
Asn	n/a	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**
Lys	n/a	AAG	AAG	**AAG**	**AAG**	**AAG**
Asp	n/a	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**
Glu	n/a	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**
Cys	n/a	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**
Trp	n/a	**UGG**	**UGG**	**UGG**	**UGG**	**UGG**
Arg	n/a	**CGU**, CGC*	CGC*	CGC*	**CGU**, CGC*	**CGU**, CGC*
Gly	n/a	GGU, **GGC**	**GGC**	**GGC**	**GGC**	**GGC**
Stop	n/a	UAA	UAA	UAA	UGA	UAA

Note. Optimal codons which complement the major tRNA gene are written in bold. An asterisk denotes an optimal codon which would complement the major tRNA gene if the adenosine at the wobble position underwent deamination modification.

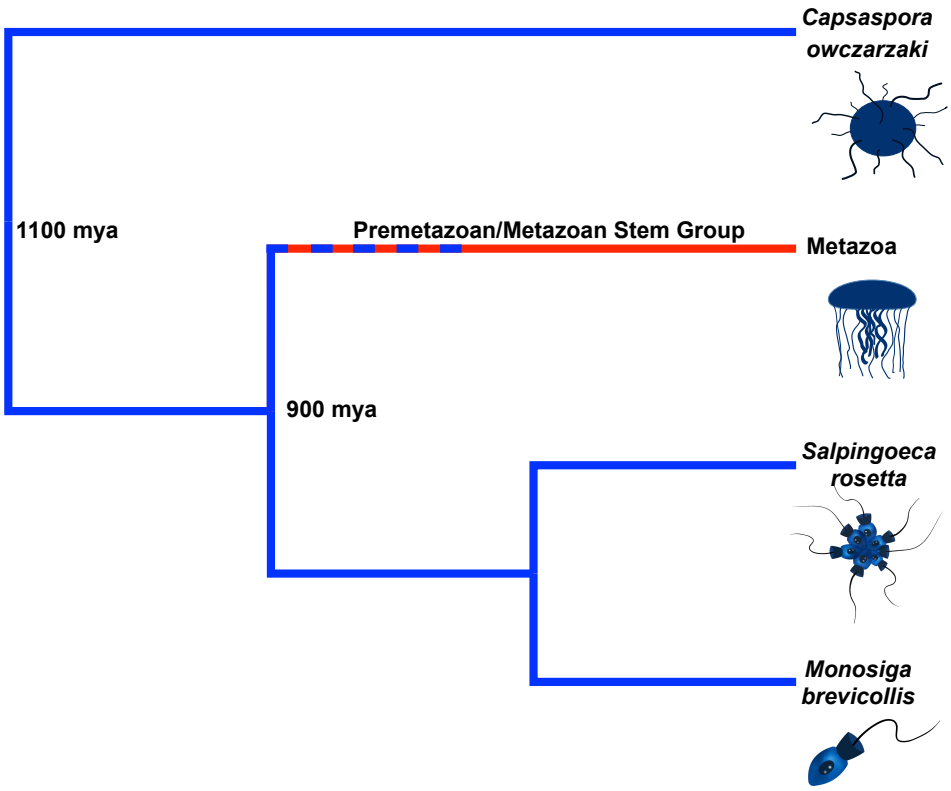
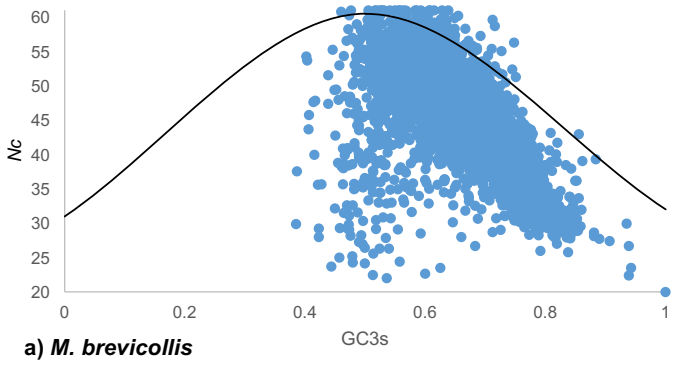
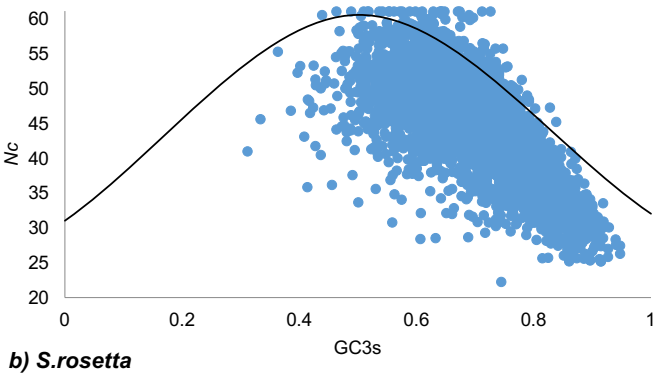

Capsaspora
owczarzaki
1100 mya
Premetazoan/Metazoan Stem Group
Metazoa
900 mya
Salpingoeca
rosetta
Monosiga
brevicollis

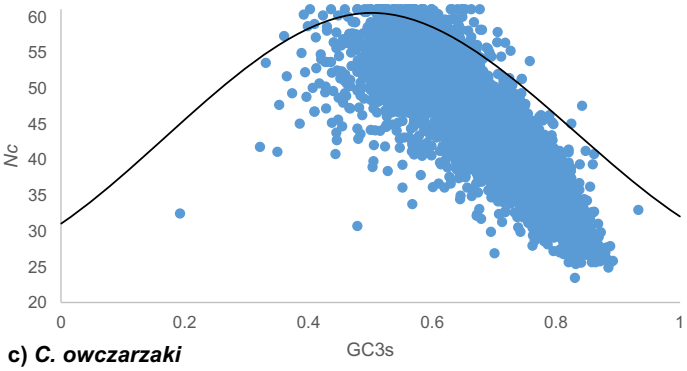
Figure 1



a) ***M. brevicollis***



b) S.rosetta



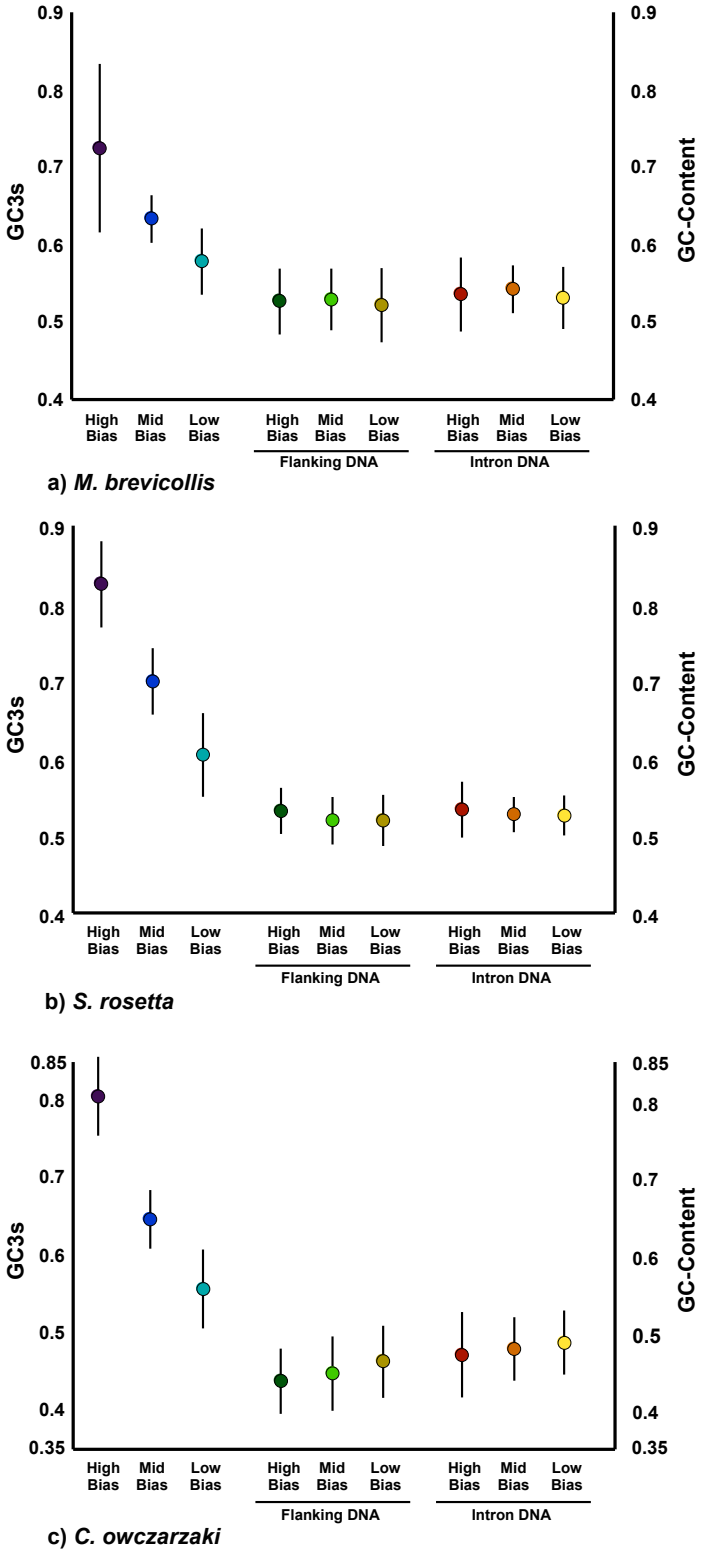
c) ***C. owczarzaki***

863

864 **Figure 2**

865

866

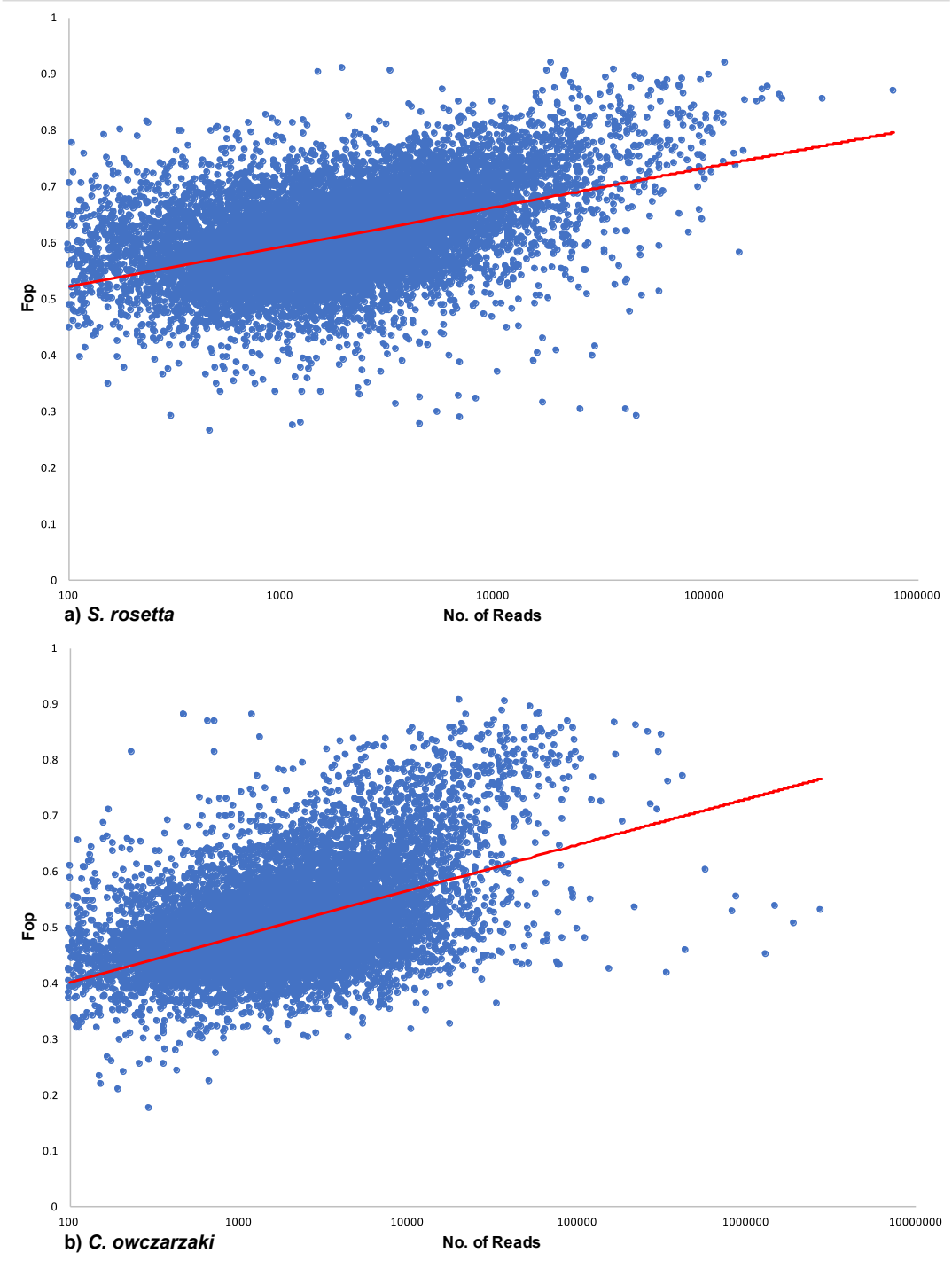


867

868 **Figure 3**

869

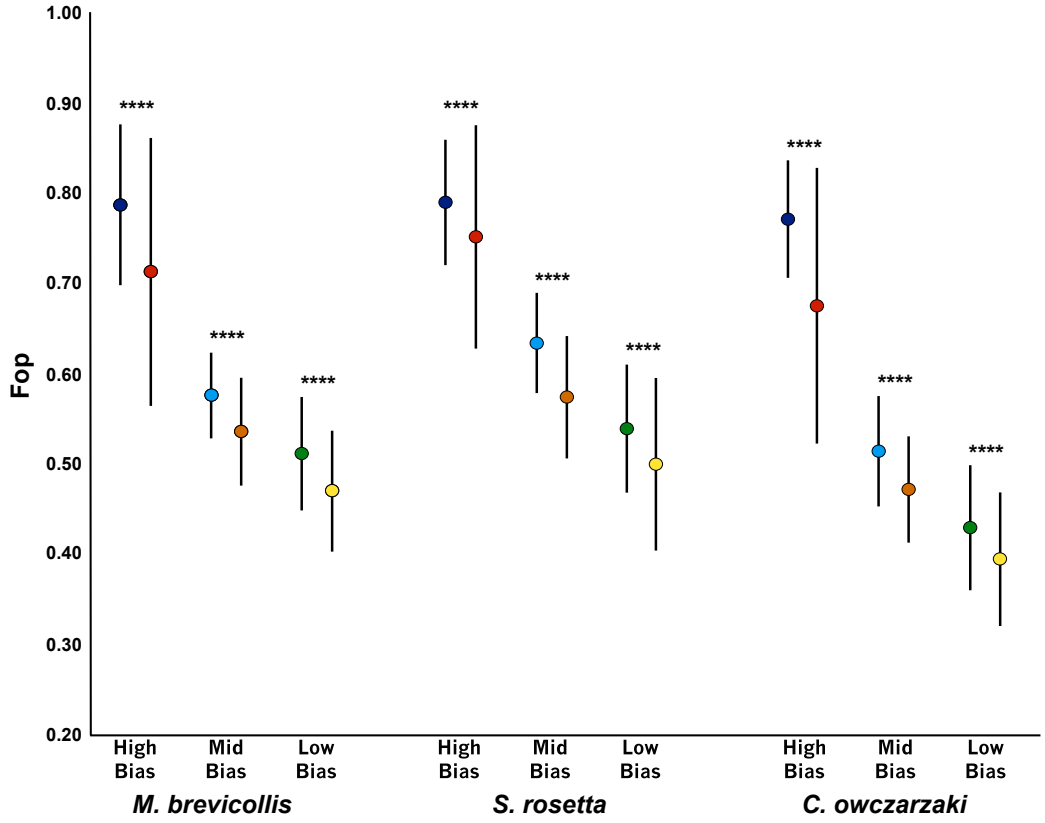
870



871

872 **Figure 4**

873



874

875 **Figure 5**

876